Genomics 91 (2008) 492–502

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

DNA methylation profiling of pseudogene–parental gene pairs and two gene families

Rene Cortese, Manuel Krispin, Gunter Weiss, Kurt Berlin, Florian Eckhardt *,[1]

Epigenomics AG, Kleine Präsidentstrasse 1, 10178 Berlin, Germany

ARTICLE INFO

Article history:
Received 30 July 2007
Accepted 1 February 2008
Available online 2 May 2008

Keywords:
DNA methylation
Gene duplication
Epigenesis, genetic
Pseudogenes

ABSTRACT

A substantial proportion of human genes contain tissue-specifically DNA-methylated regions (TDMRs). However, little is known about the evolutionary conservation of differentially methylated loci, how they evolve, and the signals that regulate them. We have studied TDMR conservation in the PLG and TBX gene families and in 32 pseudogene–parental gene pairs. Among the members of the recently evolved PLG gene family, 5′-UTR methylation is conserved and inversely correlated with the cognate gene expression, indicating as well a conserved regulatory role of DNA methylation. Conversely, many genes of the much older TBX family display complementary tissue-specific methylation, suggesting an epigenetic complementation in the evolution of this gene family. Similar to gene families, unprocessed pseudogenes arose from gene duplications and we found TDMR conservation in some pseudogene–parental gene pairs displaying short evolutionary distances. However, for the majority of unprocessed pseudogenes and for all processed pseudogenes examined, we found that tissue-specific methylation arose de novo after gene duplication.

Mammalian DNA methylation constitutes an important layer of epigenetic control and has been implicated in the control and regulation of tissue-specific gene expression, gene imprinting, X-chromosomal activation, and chromosomal integrity (reviewed in [1]). Additionally, aberrant methylation and associated tumor suppressor gene silencing and chromosomal instability have been shown in numerous neoplasias [2,3].

Several recent reports have recognized the widespread occurrence of tissue-specific differentially methylated regions (TDMRs) in healthy tissues [4–7]. TDMRs frequently map to the 5′ regions of annotated genes, but many are located in exon, introns, or intergenic regions. In addition, differential methylation is not limited to coding genes, but has been observed in pseudogenes as well [4,8]. Currently, little is known about how TDMRs evolve and are maintained and the signals that regulate them. One possibility is that differentially methylated loci might bear an intrinsic genetic or epigenetic signature that triggers differential methylation. Another possibility, not mutually exclusive, is that differential methylation is the result of other *cis*- or *trans*-acting factors or sequences that regulate the methylation profile of a particular locus. In this regard, gene duplication events represent a well-suited model to address the dynamics of (differential) methylation.

Gene duplications are a source of genomic novelties leading to gene families and unprocessed pseudogenes. Using comparative analysis, Lynch and Conery [9] estimated that gene duplications arise at a high frequency of about 0.01 per gene per million years. The classical model by Ohno [10] predicts that upon gene duplication, one copy retains the original function under strong surveillance by negative selection, while the other copy becomes free of selective constraints and evolves mainly in a neutral fashion. Hence, the most likely fate of a new gene duplication event is its mutational degradation into a pseudogene. However, gene duplication events may lead as well to genes with either a reduced functional capacity (subfunctionalization) or a new function (neofunctionalization), but little is known about the mechanisms that prevent degradation into a nonfunctional gene. In contrast to unprocessed pseudogenes, which generally maintain the exon/intron structure of their parental genes, processed pseudogenes are duplicated genes that arose due to the reverse transcription of the parental gene mRNA and typically lack both regulatory sequences and an exon/intron structure [11].

Here, we study the dynamics of DNA methylation in three gene duplication events leading to either functional gene families or unprocessed or processed pseudogenes. We selected two gene families that differ greatly in their evolutionary age, the ancient TBX family, which exists in vertebrate and invertebrates, and the more recently evolved plasminogen precursor (PLG) family, which is present only in the hominoid lineage. The unifying feature of the TBX transcription factor family is the presence of the T domain, which confers DNA binding and dimerization. In mammals, 17 distinct genes have been identified that can be grouped into five subfamilies (the T, TBX1, TBX2, TBX6, and Tbr1 subfamilies) based on their evolutionary distance [12]. TBX transcription factors are crucial in regulating a plethora of processes such as craniofacial development, limb outgrowth and patterning, and T cell differentiation [12]. The second gene family we investigated is the PLG family, consisting of four known members (*PLG, PLGLA, PLGLB1* and *PLGLB2*). While the *PLG* gene itself is located on chromosome 6q25.3 within the *IGF2R* imprinting cluster, the three plasminogen-related genes, *PLGLA*, *PLGLB1* and *PLGLB2* map to chromosomes 2q12.2, 2p11.2 and 2q11–p11, respectively [13,14]. *PLG* encodes plasminogen that circulates in the plasma as a proenzyme [15] and, in the presence of a fibrin clot, is converted to plasmin by the tissue plasminogen activator. Upon this

* Corresponding author. Fax: +49 30 24345 555.
E-mail address: f.eckhardt@brahms.de (F. Eckhardt).
[1] Present address: Brahms AG, Neuendorfstrasse 25, 16761 Hennigsdorf, Germany.

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.02.004

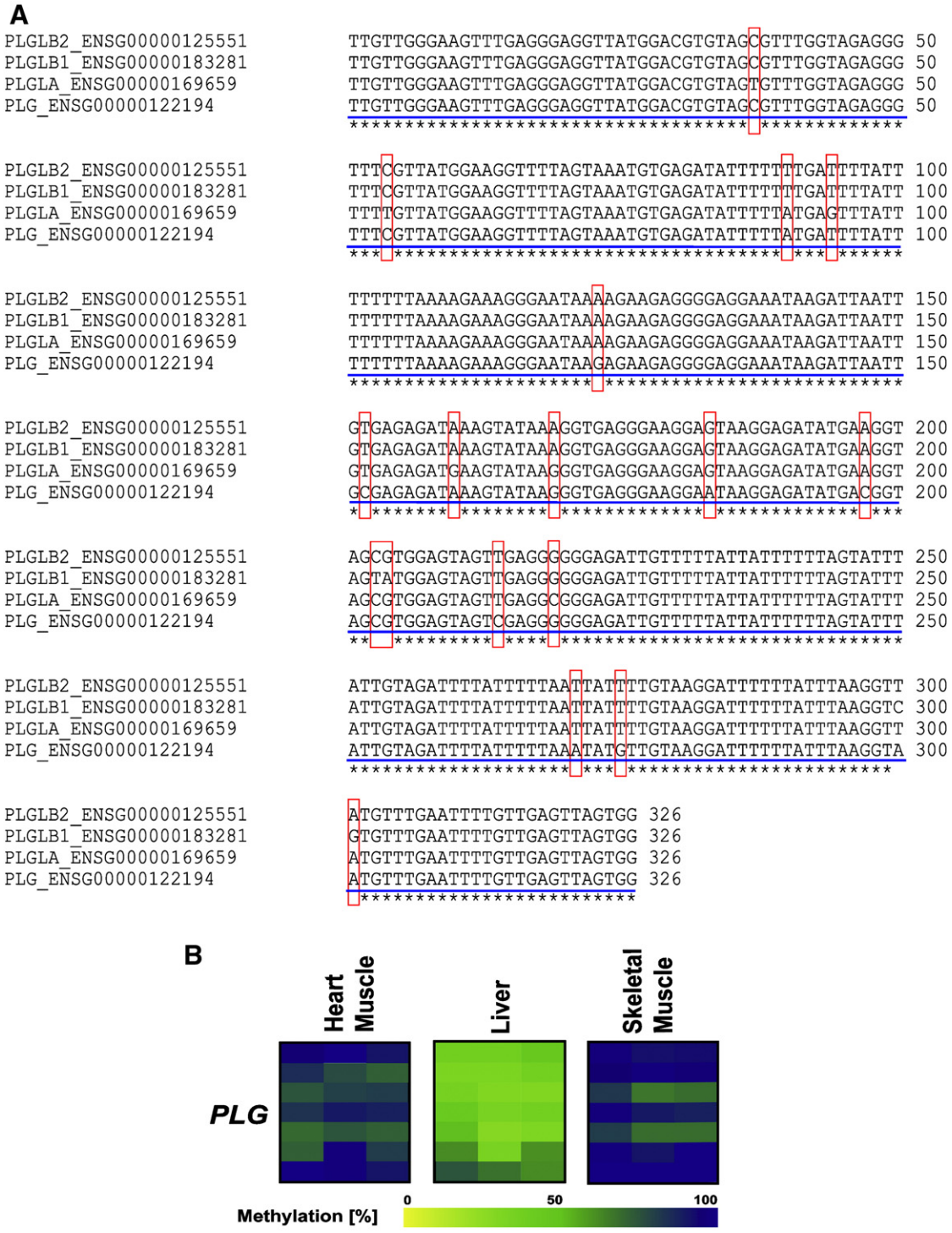

Fig. 1. Tissue-specific methylation and expression are conserved among the members of the PLG gene family. (A) *PLG* gene and its variants display a high homology of sequence as evidenced by the genomic sequence alignment of the *PLG*, *PLGLA*, *PLGLB1* and *PLGLB2* 5′ regions. Sequences analyzed by bisulfite sequencing are underlined. Red rectangles highlight the mismatches between the aligned sequences. (B) DNA methylation analysis for *PLG* in human adult skeletal muscle, heart muscle, and liver. Samples are displayed column-wise with rows representing individual CpG's of the PCR fragment. Quantitative methylation analysis results are shown in a color scale ranging from yellow (≈0% methylation), to green (≈50% methylation), to dark blue (≈100% methylation). (C) Mosaic distribution of DNA methylation in *PLG*, *PLGLA*, and *PLGLB* in liver. CpG sites in the subcloned PCR products were either all methylated or all unmethylated. Numbers indicate the position of each CpG relative to the corresponding transcription start site (TSS) (Ensembl NCBI 40). Filled and empty circles represent methylated and unmethylated CpG's, respectively. Dashed circles indicate CpG positions lost due either to sequence polymorphism (in the *PLG* gene) or to sequence mismatches with *PLG* (*PLGLA* and *PLGLB* genes). Alleles were identified by an annotated SNP (rs4252059) within the amplified sequence for *PLG* or the equivalent polymorphism in *PLGLA* and *PLGLB*. (D) Expression of *PLG* in human heart muscle, liver, and skeletal muscle. Results of RT-PCR show that gene silencing correlates with 5′-UTR hypermethylation of *PLG*. Similar amounts of cDNA were used as indicated by control amplification of actin β (*ACTB*). Representative results for three independent samples are shown. Total RNAs derived from mixed tissues and cell lines were used as positive control. (E) Biallelic expression of *PLG*. Biallelic expression was analyzed by sequencing the amplified cDNA and identification of a heterozygous annotated SNP (red arrow, SNP rs1136056). Presence of both alleles in the sequenced cDNA indicates biallelic expression.

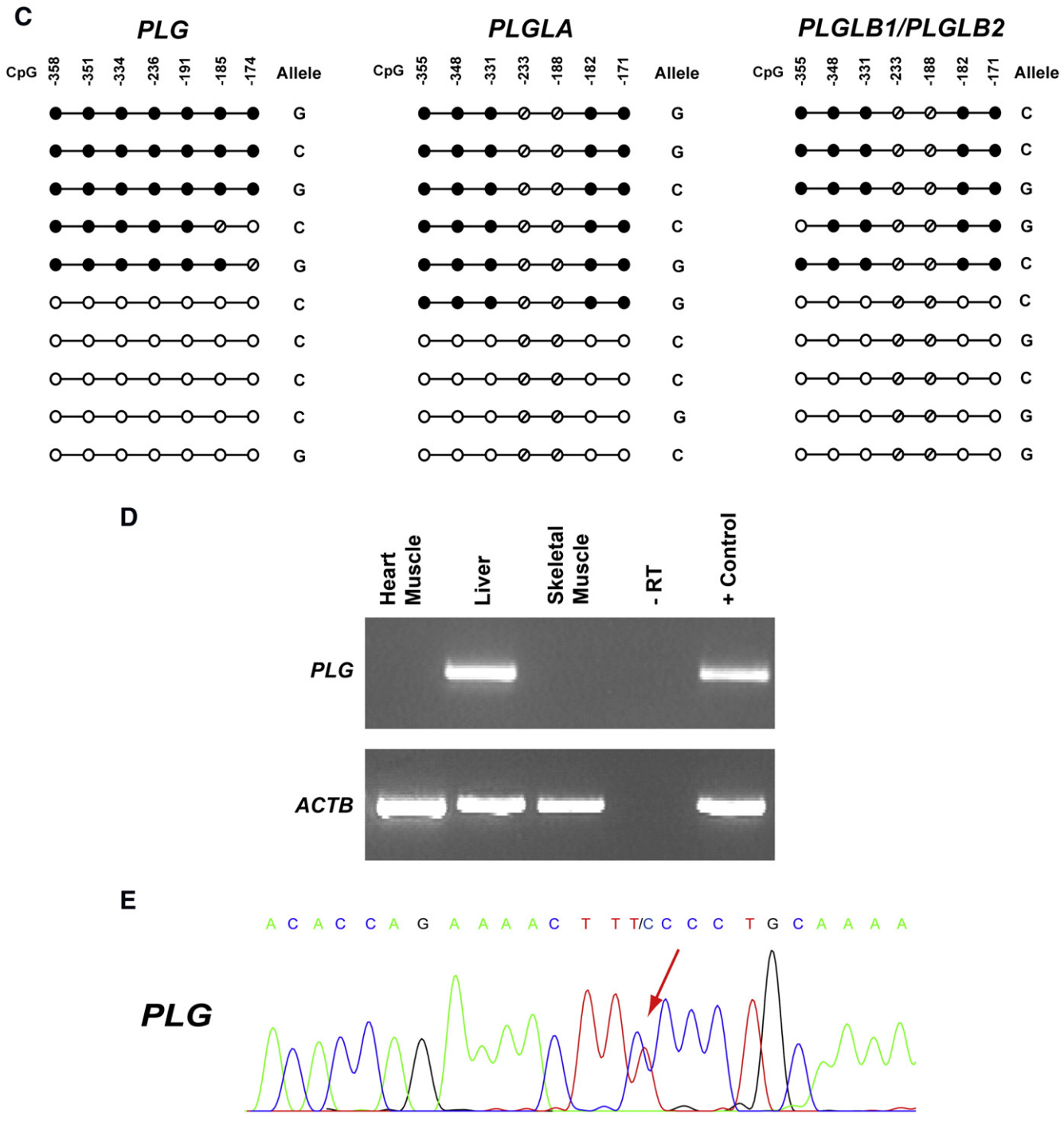

Fig. 1 (*continued*).

activation, the proteolytic plasmin digests the insoluble fibrin clot, playing a key role in tissue repair and wound healing. However, the functions of *PLGLA*, *PLGLB1* and *PLGLB2* remain largely elusive.

Results

DNA methylation and expression of PLG gene family members

As part of a chromosome-wide methylation profiling study, we recently obtained the methylation profile of *PLG* [4]. In that study, we found that the 5′-UTR of *PLG* is differentially methylated in healthy tissues, with skeletal and heart muscle being 100% methylated and liver being only 50% methylated (Fig. 1B).

Comparative sequence analysis of human *PLG* showed a high homology to three members of the *PLG* gene family (Fig. 1A). The *PLGLA* gene (ENSG00000169659) is located on chromosome 2q12.2 and displays a 95.8% homology with the *PLG* gene in its 5′ region and exon 1. Similarly, *PLGLB1* (ENSG00000125551) and *PLGLB2* (ENSG00000183281) are located on chromosome 2p11.2 and 2p11-q11, respectively, with DNA sequence homology in the 5′ region to *PLG* of about 95 and 96%, respectively.

We examined, if the differential methylation observed in the 5′-UTR of *PLG* is conserved in these genes and if the cognate mRNA expression patterns display an inverse correlation with the respective promoter methylation that would point to a conserved regulatory control of expression by DNA methylation. To this end, we used several sequence

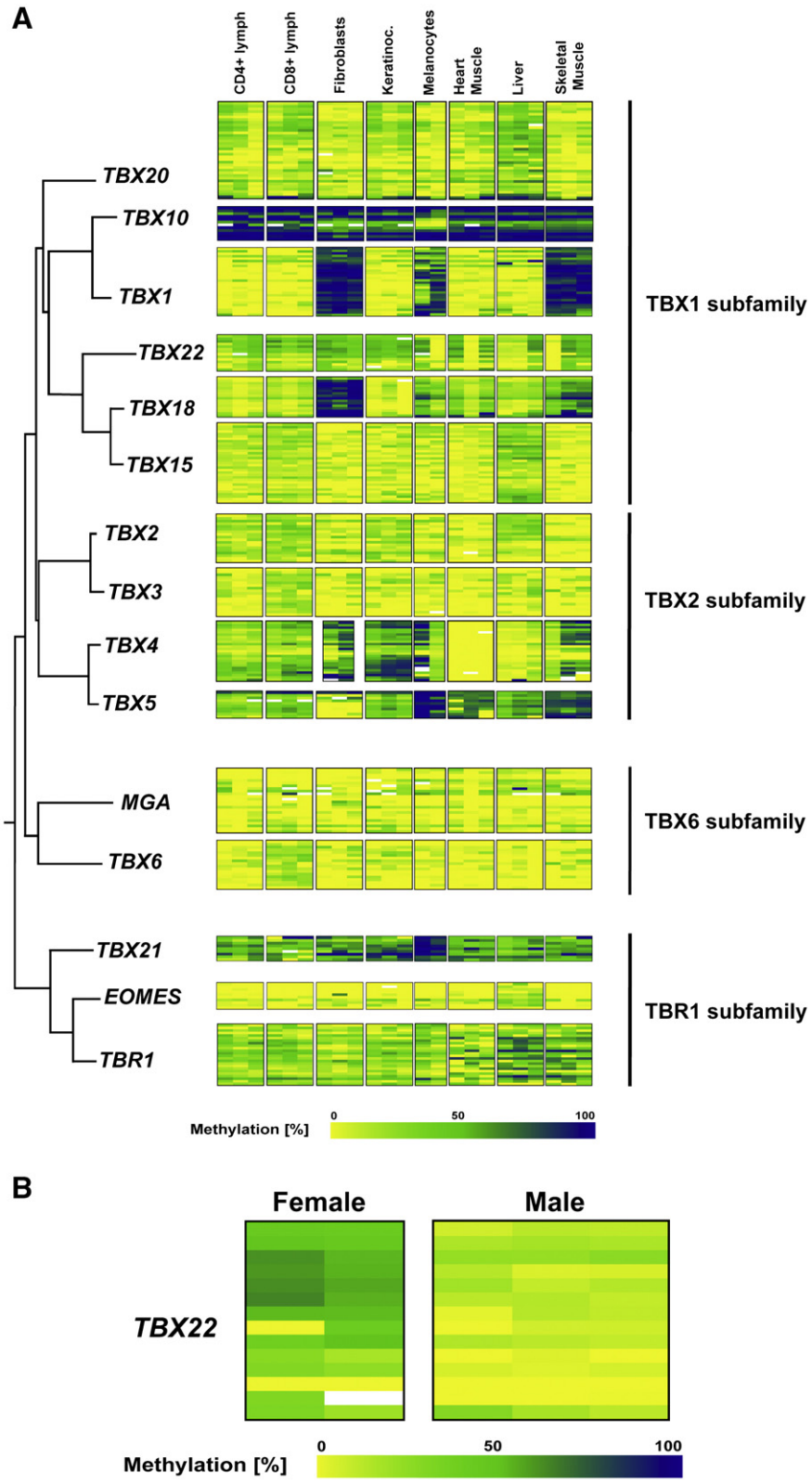

Fig. 2. Methylation profiling of TBX gene family members. (A) Methylation profiles of 15 genes of the TBX transcription factor family and their evolutionary relationships are shown. The 15 genes are from four subfamilies (Tbr1 subfamily, *TBX21, EOMES, TBR1;* Tbx1 subfamily, *TBX1, TBX10, TBX15, TBX18, TBX20,* and *TBX22;* Tbx2 subfamily, *TBX2, TBX3, TBX4,* and *TBX5;* Tbx6 subfamily, *MGA* and *TBX6*). Closely related TBX genes display complementary methylation profiles (e.g., *TBX1*/*TBX10* and *TBX2*/*TBX3*). The evolutionarily more distant *MGA* and *TBR1* are not differentially methylated in the tissues examined. The methylation values are color coded as in Fig. 1. Phylogenetic relationships within the TBX family were previously described in [34]. (B) Gender-specific methylation in the *TBX22* gene located on chromosome X. *TBX22* is approx 50% methylated in liver samples derived from female donors and approx 0% in male-derived liver samples. The methylation values are color coded as in Fig. 1.

Table 1
Methylation values for pseudogenes and their respective parental genes

Pseudogene	Sample 1 methylation[a]	Sample 2 methylation	Parental gene	Sample 1 methylation	Sample 2 methylation	Sequence homology	Type pseudogene
AP000357.2	100%	0%	ACTR2	100%	100%	57.10%	Processed
AP000357.3	100%	25%	ARL5A	0%	0%	92.30%	Processed
RP11-758C21.1	100%	50%	BDH2	100%	100%	96.60%	Processed
RP1-181C9.1	100%	25%	ANP32B	0%	0%	71.40%	Processed
AP000358.2	100%	25%	FN3K	100%	50%	87.70%	Processed
CTA-229A8.2	100%	50%	GAPDH	50%	50%	62.50%	Processed
AC007050.7	100%	100%	POM121L3	100%	100%	96.80%	Processed
RP3-412A9.4	100%	50%	SNRPN	100%	100%	88.10%	Processed
KB-1269D1.3	100%	25%	MAD1L1	100%	75%	95.10%	Processed
AC000078.2	100%	25%	RPL8	100%	75%	65.90%	Processed
AC004019.3	100%	50%	LOC376522	75%	50%	84.00%	Processed
CTA-373H7.4	100%	0%	HBLD1	0%	0%	99.50%	Processed
CTA-373H7.4	100%	0%	HBLD1	75%	75%	75.80%	Processed
RP1-47A17.8	100%	25%	ADAMTS7	100%	100%	84.50%	Processed
RP1-106I20.2	100%	100%	NDUFA9	100%	100%	94.10%	Processed
AC004471.4	100%	100%	NM_032028.2	100%	100%	82.80%	Processed
RP3-405J24.1	100%	100%	RPL12	100%	50%	92.20%	Processed
CTA-150C2.8	100%	50%	APOBEC3G	100%	100%	64.30%	Unprocessed
CTA-246H3.2	0%	0%	LRP5	0%	0%	89.10%	Unprocessed
RP4-539M6.7	100%	25%	SLC39A1	100%	25%	84.80%	Unprocessed
KB-1592A4.6	100%	100%	BCR	100%	100%	97.20%	Unprocessed
LL22NC03-31F3.7	100%	50%	Q6UW61_HUMAN	100%	50%	94.40%	Unprocessed
RP11-34P13.1	100%	100%	DDX11	75%	75%	71.30%	Unprocessed
RP11-223J15.2	100%	100%	EEF1A2	50%	50%	64.30%	Unprocessed
RP4-732G19.2	100%	100%	CYP4Z1	100%	100%	88.80%	Unprocessed
RP11-552J9.1	100%	100%	XAGE2	100%	100%	72.60%	Unprocessed
BMS1LP6	100%	100%	BMS1L	100%	100%	89.10%	Unprocessed
CTSLL3	75%	75%	CTSL	75%	75%	89.30%	Unprocessed
RP11-432I13.3	75%	100%	CUBN	100%	75%	89.40%	Unprocessed
RP11-453N3.6	75%	75%	ABCD1	75%	75%	95.40%	Unprocessed
RP11-392A23.3	50%	50%	GSTA1	100%	100%	83.70%	Unprocessed
NM_002688.4	0%	100%	SEPT5	0%	0%	53.50%	Unprocessed

[a]DNA methylation values corresponding to two paired samples, being tissues (liver, skeletal muscle, and heart muscle) or primary cell lines (keratinocytes, fibroblasts, melanocytes, and $CD4^+$ lymphocytes), per parental gene–pseudogene pair are shown. The selection was based on pseudogenes, for which we had previously detected tissue-specific methylation [4]. The values shown are the medians of all CpG positions within the PCR fragment rounded to 0, 25, 50, 75, and 100%.

mismatches of the PLG isoforms to infer the methylation status of each gene and the identity of the expressed transcripts. However, because *PLGLB1* and *PLGLB2* differ by only one mismatch in the coding region (data not shown), we could not assess differences in methylation and mRNA expression between these two genes.

We analyzed the conservation of DNA methylation in human *PLG* and the *PLGLA* and *PLGLB1/B2* genes by sequencing of subcloned PCR amplicons derived from bisulfite-treated DNA obtained from healthy adult skeletal muscle, heart muscle, and liver. This approach revealed that in liver samples, the DNA methylation of all PLG genes (*PLG, PLGLA,* and *PLGLB1/B2*) is distributed in a mosaic manner (Figs. 1B and C), with all CpG's contained in the same clone being either methylated or unmethylated. In contrast, DNA derived from heart muscle or skeletal muscle displayed a homogeneous methylation for all clones (Fig. 1B and data not shown). Methylation of *PLG* clones did not segregate with an annotated SNP in this region (rs4252059), indicating that this gene is not allele-specifically methylated. Similarly, we found heterozygous polymorphisms in the sequences corresponding to *PLGLA* and *PLGLB* subclones, pointing to a biallelic methylation of both variants.

Expression analysis by RT-PCR for *PLG* revealed that the 5′-UTR methylation of this gene is inversely correlated with its expression (Fig. 1D). In tissues (heart muscle and skeletal muscle) that displayed a completely methylated *PLG* 5′-UTR we detected no *PLG* mRNA expression, while in liver, the respective 5′-UTR was only 50% methylated and *PLG* was expressed. Subsequent sequencing of the obtained RT-PCR product revealed the presence of a heterozygous SNP (rs1136056), indicating that *PLG,* similar to the biallelic methylation of its 5′-UTR, is biallelically expressed as well (Fig. 1E). To study the expression of the other *PLG* genes, we designed a RT-PCR fragment that allowed for the unbiased amplification of the three transcripts. Primers for this fragment bound to regions containing no mismatches between *PLG, PLGLA,* and *PLGLB* transcripts, whereas the amplified region contained nine mismatches that allowed the identification of the expressed gene. Similar to *PLG* (Fig. 1E), *PLGLA* and *PLGLB1/PLGLB2* were expressed only in liver, not in skeletal or heart muscle (data not shown), indicating a DNA methylation-dependent regulation similar to that of *PLG*.

Differential DNA methylation in the TBX family

In contrast to the *PLG* gene duplicates that arose recently in the hominoid lineage, the *TBX* family arose early during evolution and is present in vertebrates, invertebrates, and protostomes (e.g., *Drosophila melanogaster* and *Caenorhabditis elegans*). To examine further the methylation profiles in this large gene family we selected 15 genes from four subfamilies (Tbr1 subfamily, *TBX21, EOMES,* and *TBR1;* Tbx1 subfamily, *TBX1, TBX10, TBX22, TBX18, TBX15,* and *TBX20;* Tbx2 subfamily, *TBX2, TBX3, TBX4,* and *TBX5;* Tbx6 subfamily, *MGA* and *TBX6*) and performed DNA methylation profiling of their 5′-UTRs in eight different tissues and primary cells (Fig. 2A). Most of the genes and tissues were unmethylated, with *TBX10* being the most prominent exception. Of these genes, 7 (*TBX10, TBX1, TBX18, TBX15, TBX4, TBX5,* and *TBX21*) were differentially methylated in at least one of the tissues, but none of the genes showed an identical methylation pattern (Fig. 2A). In particular, differential methylation was evident in genes displaying a shorter evolutionary distance. For example, *TBX1* was hypermethylated in melanocytes, fibroblasts, and skeletal muscle, while being unmethylated in the remaining tissues. The closely related *TBX10* gene was hypermethylated in most of the tissues, but four CpG's, 93, 104, 119, and 139 bp upstream from the TSS, were unmethylated. Similarly, genes of the Tbx2 subfamily were differentially methylated as well. For *TBX22,* which is located on the X chromosome, we observed no tissue-specific methylation, but did

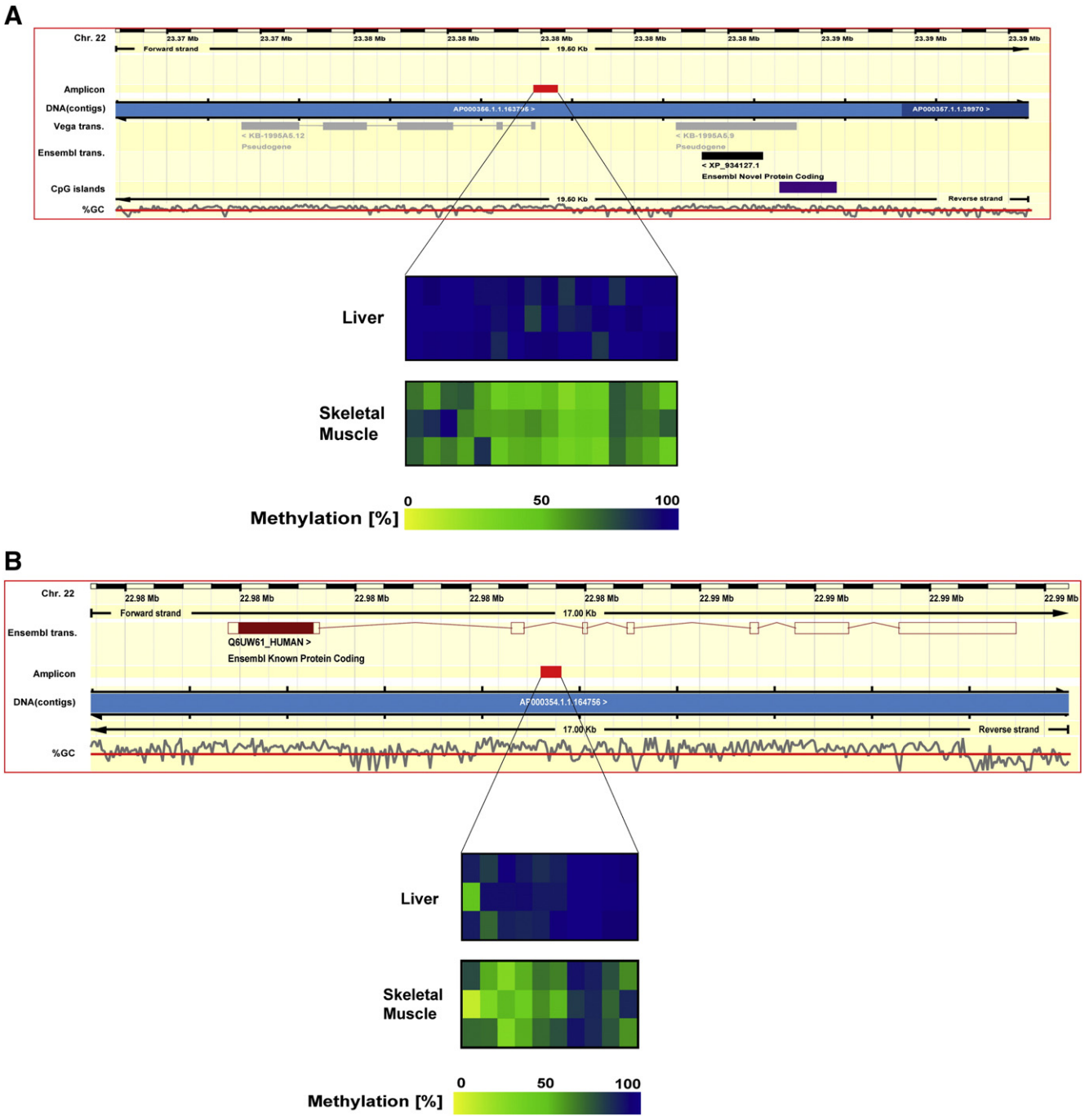

Fig. 3. Conservation of tissue-specific methylation in an unprocessed pseudogene–parental gene pair. (A and B) The TDMR of the unprocessed pseudogene *KB-1995A5.12* (A) is conserved in its parental gene *Q6UW61_HUMAN* (B). The positions of the analyzed regions are shown. Liver samples are hypermethylated in both regions, while skeletal muscle samples display a 50% methylation. Rows represent individual samples with columns representing individual CpG's of the PCR fragment. Quantitative methylation analysis results are shown in a color scale ranging from yellow (≈0% methylation), to green (≈50% methylation), to dark blue (≈100% methylation).

observe sex-specific methylation, with approx 50% methylation in samples derived from female donors and approx 0% methylation in male-derived samples (Fig. 2B). Genes such as *MGA* and *TBR1,* that are more distant from other family members, displayed no differential methylation.

DNA methylation in pseudogenes

To study further the fate of DNA methylation upon gene duplication, we examined the methylation profiles of processed and unprocessed pseudogenes. We selected putative parental gene–pseudogene pairs according to their annotation in the Vertebrate Genome Annotation Database (Vega) [16] and studied the respective DNA methylation in regions of highest sequence homology. In total, we selected 17 processed and 15 unprocessed pseudogenes for which we had previously obtained methylation profiles [4] and analyzed the DNA methylation of their respective parental genes. Respective pseudogenes were selected if they displayed tissue-specific differential methylation in at least one tissue analyzed. As controls, we included pseudogenes showing homogeneous hypermethylation (>80%) or unmethylation (<20%) in the analyzed tissues. For each pseudogene–parental gene pair, we obtained methylation profiles

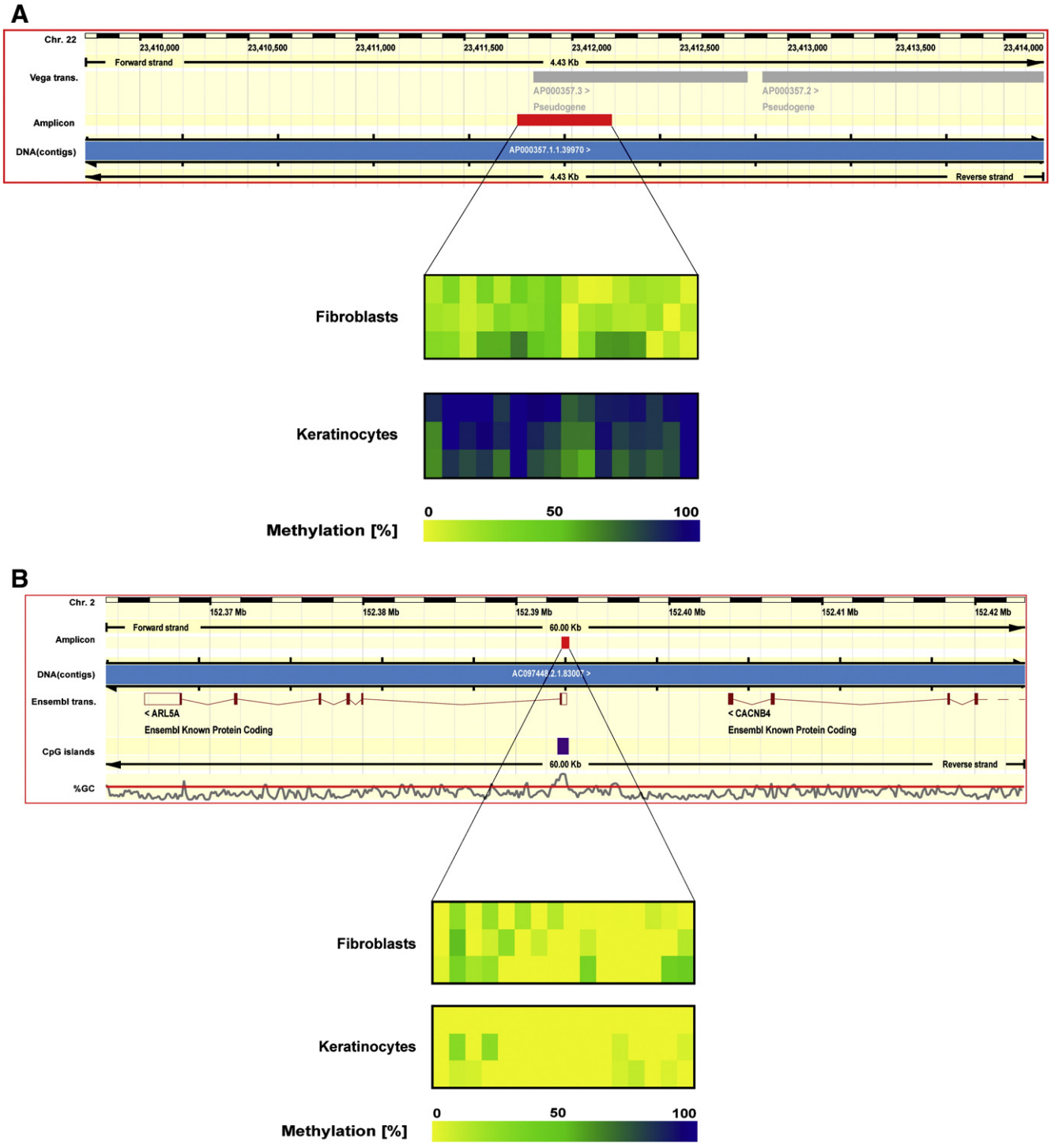

Fig. 4. Tissue-specific methylation is not conserved in processed pseudogene–parental gene pairs. (A and B) The TDMR located within the processed pseudogene *AP000357.3* (A) is not conserved in its parental gene *ARL5A* (B). The processed pseudogene is unmethylated (<20%) in fibroblasts and hypermethylated (>80%) in keratinocytes. In contrast, the parental gene is unmethylated in both cell types. Methylation values are shown as in Fig. 3.

from the same two tissues (the tissues studied included liver, skeletal muscle, and heart muscle) or primary cell lines (keratinocytes, fibroblasts, melanocytes, $CD4^+$ lymphocytes). Table 1 summarizes the observed methylation values and the homology between the loci used to measure DNA methylation of pseudogene–parental gene pairs.

Unprocessed pseudogenes

Among the 15 unprocessed pseudogenes analyzed, 5 (33%) showed tissue-specific methylation in at least one examined tissue. For 2 pseudogenes, *RP4-539 M6.7* and *LL22NC03-31F3.1,* this differential methylation was conserved in the respective parental genes, *SLC39A1* and *Q6UW61_HUMAN.* Six (40%) unprocessed pseudogenes and their respective parental genes were hypermethylated in both tissues, while 1 (7%) pseudogene–parental gene pair was unmethylated in the analyzed tissues. The remaining 3 unprocessed pseudogenes (20%) and their respective parental genes displayed heterogeneous methylation values ranging from 25 to 75%. In no case did we find a parental gene displaying tissue-specific methylation, while the respective unprocessed pseudogene had lost the differential

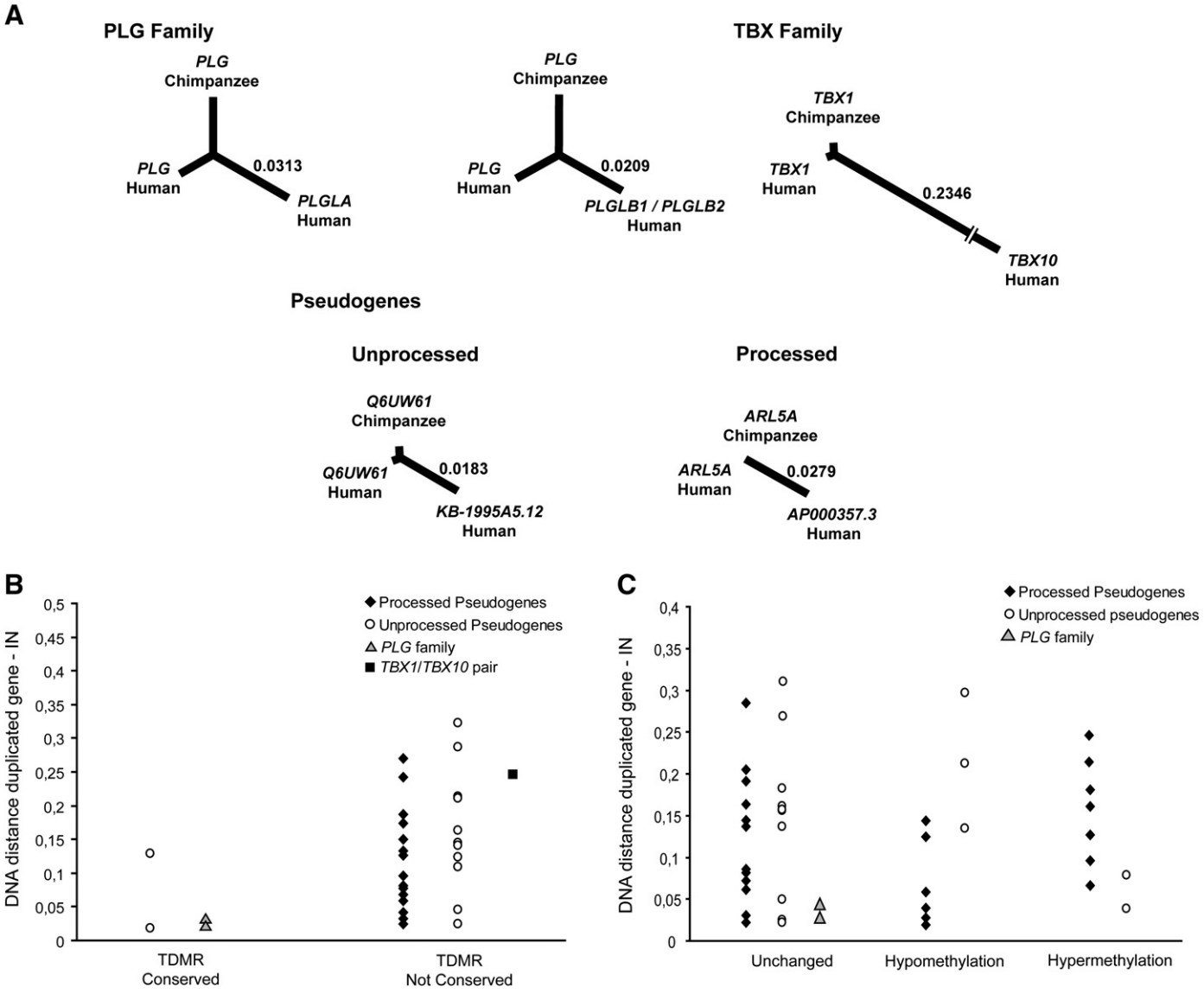

Fig. 5. DNA distances and DNA methylation in duplicated sequences. (A) Examples of DNA distances between gene variants and pseudogene–parental gene pairs. DNA distances were calculated using the PHYLIP software and results are represented as a radial tree for each gene pair. The distance between the internal node and the duplicated gene or pseudogene is indicative of the evolutionary distance to the parental gene. Examples of distances from the internal node are shown for the PLG variants (0.0313 and 0.0209, for *PLGLA* and *PLGLB*, respectively) and the *TBX1*/*TBX10* pair (top) and unprocessed (0.0183) and processed pseudogenes (0.0856; bottom). The DNA distance observed for *TBX1*/*TBX10* (0.2346) is about 10 times larger than those observed in the PLG family. (B) Relationship between tissue-specific methylation conservation and DNA distance in gene families and unprocessed and processed pseudogenes. For the TBX family, the representative *TBX1*/*TBX10* pair is shown. Pairs displaying conservation of methylation are among those showing the shorter distances. Since some parental gene–pseudogene pairs displayed identical DNA distances, only 13 points (of 15) corresponding to unprocessed pseudogenes and only 15 (of 17) processed pseudogenes are shown. (C) DNA methylation transitions in gene duplications. The methylation profiles from the parental genes were compared to the respective gene variants and pseudogenes. If the methylation differed by less than 20% in the same tissue, the methylation of the respective parental gene/pseudogene pair was considered to be unchanged. Many of the parental genes displayed the same methylation values compared to their respective pseudogenes in each individual tissue. Hypermethylated processed pseudogenes exhibited larger DNA distances to the parental genes than unmethylated processed genes. *y* axis: DNA distances to the internal node calculated by PHYLIP.

methylation. Fig. 3 illustrates the conservation of TDMRs of an unprocessed pseudogene (*KB-1995A5.12*, Fig. 3A) and its corresponding parental gene (*Q6UW61_HUMAN*, Fig. 3B). The TDMR located in the 5′ region of this unprocessed pseudogene displayed a very high sequence homology (>94%) with a region in intron 2 of the parental gene. This high sequence homology extended to other regions of this gene pair, e.g., exon 3 of *Q6UW61_HUMAN* was similarly homologous to exon 1 of *KB-1995A5.12*, suggesting that some truncation occurred during or after the duplication event. Similarly, analysis of the 5′-UTR of the unprocessed pseudogene *RP4-539M6.7* and the respective homologous region of its parental gene *SLC39A1* displayed as well a conservation of the tissue-specific methylation (Table 1 and data not shown). Expression analysis by RT-PCR for these two parental gene–pseudogene pairs revealed no correlation between the methylation status of a gene and its expression. The parental gene *Q6UW61_HUMAN* and its unprocessed pseudogene were expressed in all examined tissues. In contrast, *SLC39A1* was expressed in all the studied tissues, while no transcript was found for its unprocessed pseudogene (data not shown).

Processed pseudogenes

Among the 17 analyzed processed pseudogenes, 13 (76%) displayed tissue-specific methylation in at least one of the tissues analyzed, while 4 (24%) were hypermethylated in all tissues (Table 1). In contrast to the unprocessed pseudogenes, we did not find a conservation of tissue-specific methylation between any of the analyzed processed pseudogene–

parental gene pairs. Fig. 4 shows an example of the tissue-specific methylation of a processed pseudogene (*AP000357.3*, Fig. 4A) and the methylation profile of the respective homologous region in the parental gene *ARL5A* (Fig. 4B). The TDMR of the pseudogene displayed an average methylation in dermal fibroblasts and keratinocytes of 25 and 90%, respectively. Such differential methylation was not observed for the parental gene *ARL5A* (Fig. 4B), despite the high sequence homology of both analyzed sequences (92%). Four processed pseudogenes displayed homogeneous hypermethylation in the analyzed tissues. Among them, three of the respective parental genes were similarly hypermethylated in the same samples, *POM121L3*, *NDUFA9*, and *NM_032028.2*, corresponding to *AC007050.7*, *RP1-106I20.2*, and *AC004471.4*, respectively (Table 1). For only one gene (*RPL12*) did we find a TDMR in the parental gene that was lost in its cognate processed pseudogene, *RP3-405J24*, being hypermethylated in both tissues.

Comparative analysis

To analyze further the evolutionary relationship between the pseudogene–parental gene pairs and its relation to the observed methylation profiles, we calculated their evolutionary distance by assuming that each base in the DNA sequence has an equal chance of mutating (Jukes-Cantor model [17]). Using the chimpanzee orthologue of the parental gene as an outgroup, DNA distance matrices were computed for each pair. To minimize any bias introduced by the selection of promoter vs coding regions we used only putative coding regions to calculate the distance and included only confirmed human–chimpanzee orthologues (as retrieved from the Ensembl database). Fig. 5A shows examples of the calculated DNA distances for the *PLG* family, the *TBX1*/*TBX10* gene pair, and pairs of parental genes with processed or unprocessed pseudogenes, respectively. DNA distances between the *PLG* gene and other *PLG* family genes showed similar distances for *PLGLA* (0.0313) and *PLGLB1/B2* (0.0209). As *PLGLB* orthologues exist in chimpanzee (*Pan troglodytes*) but not in Old World monkey (*Macaca mulatta*) and within the Eutheria infraclass, we estimate that the *PLG* duplication occurred after the hominid–cercopithecoid divergence, some 29–35 million years ago [18]. In contrast, the evolutionary distance of the *TBX1* and *TBX10* pair (0.2346) is about 10 times larger than that observed for *PLG* and these genes are present as well in other mammals, such as mice, indicating that this duplication occurred at least 80–130 million years ago [19].

Next, we analyzed the conservation of DNA methylation in each tissue and its dependency on the evolutionary distance for the *PLG* gene family, the *TBX1/TBX10* pair, and the pseudogene–parental gene pairs (Fig. 5B). In this analysis, conserved methylation was assigned if the methylation values from the respective gene pair differed by less than 20% between measurements of the same tissues. For both, unprocessed and processed pseudogenes, we observed a lack of methylation conservation for most of the tissues and pseudogenes examined. This lack of conserved methylation was evident as well for pseudogene–parental gene pairs that arose fairly recently as evidenced by a short evolutionary distance (Fig. 5B). Conserved TDMRs were observed for only two unprocessed pseudogene–parental gene pairs (see above) and for the *PLG* gene family. Notably, these genes displayed a rather short evolutionary distance of 0.1258 for the *RP4-539M6.7–SLC39A1* and 0.0183 for the *LL22NC03-31F3.7–Q6UW61_HUMAN* pair. As well, we analyzed if pseudogenes are preferentially hypermethylated or hypomethylated compared to their parental genes but found no evidence that the analyzed pseudogenes became persistently hypermethylated (Fig. 5C). We found pseudogenes that were both hypomethylated and hypermethylated compared to the cognate parental gene. Although the analyzed number of gene pairs is too small to generalize our observation, we observed a trend indicating that hypermethylated processed pseudogenes were evolutionarily more distant from their parental genes than unmethylated processed pseudogenes.

Discussion

The mechanisms and signals that lead to tissue-specific methylation are currently not known. Here, we have analyzed the fate of TDMRs in gene duplication events. For the functional *PLG* gene family, the TDMRs are conserved in all known members, independent of the genomic location of each gene. For all *PLG* genes, the methylation status of the TDMR is inversely correlated with the respective gene expression, suggesting that the functions of these TDMRs are conserved as well. Similarly, we found conserved TDMRs in two gene duplication events leading to unprocessed pseudogenes. Although more gene families will have to be tested to generalize our observation, these results suggest that the sequence itself might contain the signal conferring tissue-specific methylation independent of the genomic location. A possibility we currently cannot rule out is that the differential methylation of, e.g., *PLGLA* and *PLGLB1/B2* arose independently and reoccurred after gene duplication.

The TDMRs observed in the TBX family, and the vast majority of TDMRs found in both unprocessed and processed pseudogenes, are not conserved, suggesting that other mechanisms leading to TDMRs must exist as well. Possibly, once the gene duplication is transmitted through the germ line, DNA methylation associated with, e.g., tissue development may override the existing methylation mark of the duplicated gene and thereby generate a new methylation profile for this gene. We have observed a bias for recently evolved processed pseudogenes being rather unmethylated compared to more distant processed pseudogenes. This finding, if confirmed by a larger gene panel, may indicate that newly processed pseudogenes become preferentially integrated into open chromatin structures that are generally associated with unmethylated DNA. A similar bias has been reported for retroviruses such as the human immunodeficiency virus that preferentially integrate in open chromatin structures [20].

The function of TDMRs located within pseudogenes is not known. They could be nonfunctional evolutionary relics of abortive gene duplications but may as well confer stage- and tissue-specific expression of pseudogenes. Some pseudogenes, although not coding for a functional protein, are transcribed and have a regulatory function [21]. A prominent example is the *Xist* gene, a key regulator for X-chromosomal inactivation that arose by pseudogenization of a protein-coding gene [22]. Other pseudogenes, such as the *NOS* [23] and the *Markorin* [24] pseudogenes, regulate expression of their respective parental genes, although some results have been disputed by others [25].

Rodin and Riggs [26] proposed an epigenetic complementation model to predict the fate of gene duplicates. In this model, stage- and tissue-specific epigenetic silencing/activation helps to maintain negative selection on both copies of the duplicated gene. This epigenetic complementation would thus lead to a complementary expression of the original gene and its twin copy, and consequently, complementary gene expression would expose both copies to a purifying selection and may prevent degradation into a pseudogene. In this aspect, it is of interest to note that the methylation profiles observed for the functional TBX gene family were very gene-specific and each gene had a distinct methylation profile. In this study, we have analyzed a limited number of different tissues and cells and it is likely that a more comprehensive analysis would reveal further specific TDMR profiles. For example, we have recently shown that *TBX21* is specifically unmethylated in $CD4^+$ Th1 and $CD8^+$ memory lymphocytes but not in $CD4^+$ and $CD8^+$ naïve lymphocytes [27]. In contrast, the very recently evolved PLG family members displayed similar expression and methylation profiles that may indicate that some *PLG* paralogues are destined to pseudogenization. Alternatively, if *PLGLA*, *PLGLB1* and *PLGLB2* are functionally different compared to *PLG*, it is possible that these genes escaped pseudogenization by a neofunctionalization event and may share the epigenetic regulatory mechanism due to their high sequence homology. More comprehensive DNA methylation profiling studies

are needed to understand further the dynamics of DNA methylation in genome evolution and gene duplications.

Materials and methods

Tissue samples

DNA used for methylation analysis was isolated from human tissue samples (human heart, liver, and skeletal muscle) and primary cell cultures (melanocytes, fibroblasts, keratinocytes, and $CD4^+$ lymphocytes). Human tissue samples were acquired from commercial suppliers: Asterand (Detroit, MI, USA), Pathlore Plc. (Nottingham, UK), Tissue Transformation Technologies (T-Cubed, Edison, NJ, USA), Northwest Andrology (Missoula, MT, USA), NDRI (Philadelphia, PA, USA), and BioCat GmbH (Heidelberg, Germany). Primary cells were purchased from Cascade Biologics (Mansfield, UK), Cell Applications (San Diego, CA, USA), Analytical Biological Services (Wilmington, DE, USA), Cambrex Bio Science (Verviers, Belgium), and DIGZ (Berlin, Germany) and were cultured for a maximum of three passages according to the supplier's recommendations. For correlative studies, matched total RNA/DNA samples (heart muscle, skeletal muscle, liver) were purchased from BioCat GmbH. In all cases, only anonymous samples were used and ethical approval was obtained for the study. Mouse DNA from liver, heart, and skeletal muscle was acquired from BioCat GmbH.

DNA extraction and PCR amplification

DNA was extracted using the DNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's recommendations. For DNA methylation analysis, DNA was bisulfite converted and PCR amplified as previously described [28]. Bisulfite-specific primers with a minimum length of 18 bp were designed using a modified Primer3 program. The target sequence of the designed primers contained no CpG's, allowing an unbiased amplification of both hypomethylated and hypermethylated DNAs. Primers were also tested for specificity by electronic PCR. Methylation profiles of the 5′-UTRs of *PLG*, *PLGLA*, and *PLGLB1/B2* were analyzed by unbiased, simultaneous amplification of these genes. Subsequent subcloning of the PCR fragment and sequencing allowed the identification of the specific gene by the identification of several sequence mismatches (Fig. 2A). Sequence mismatches were confirmed by sequencing of genomic DNA as well.

Genomic DNA amplification was carried out using the HotStartTaq DNA polymerase kit (Qiagen) with 10 ng of genomic DNA and gene-specific primers. Primers for genomic DNA amplification were designed using the Primer3 software [29]. Amplification conditions for the genomic DNA were 15 min at 95 °C followed by 40 cycles of 92 °C for 60 s, 72 °C for 60 s, and 72 °C for 60 s and a final extension step of 10 min at 72 °C. Genomic PCR fragments used for the genotyping of the matched DNA/RNA samples contained at least one reported SNP.

RNA extraction and RT-PCR

In all cases, tissue and cell samples were kept at -80 °C prior to RNA isolation and isolated RNAs were stored at -80 °C until further use. Total RNA was isolated with the RNeasy kit (Qiagen) followed by cDNA synthesis using the Omniscript RT kit from the same supplier and random hexamers. PCR (92 °C for 1 min, 60 °C for 1 min, 72 °C for 1 min for 40 cycles) was performed using the HotStartTaq DNA polymerase kit (Qiagen) with 3 µl of the prepared cDNA and gene-specific primers. All kits were used according to the manufacturer's recommendations. PCR products were analyzed by electrophoresis on 2.5% agarose gels. All RT-PCR fragments were designed spanning at least one intron to avoid amplification of contaminating genomic DNA. Universal RNA (BioCat) was used as positive control. RT-PCR fragments used to determine allelic expression contained at least one reported SNP.

As for the bisulfite-treated DNA, a unique primer pair was used to study the expression of *PLG*, *PLGLA*, and *PLGLB1/B2*. Primers bound to sequences showing no mismatches between the transcripts, while the amplified region contained several mismatches that allowed their identification after direct sequencing.

Sequencing

PCR amplicons from bisulfite-treated and genomic DNA, as well as RT-PCR products, were quality controlled by agarose gel electrophoresis, purified with ExoSAP-IT (USB Corp., Cleveland, OH, USA) to remove any excess nucleotides and primers, and sequenced directly in forward and reverse directions. Alternatively, the resultant PCR and RT-PCR products were cloned into a TA-cloning plasmid according to the manufacturer's instruction (pGEM T-Easy Cloning Kit; Promega, Madison, WI, USA) and DNA was isolated using a Qiaprep Spin Plasmid Miniprep Kit (Qiagen) according to the manufacturer's instructions. All PCR fragments and plasmids were sequenced in forward and reverse directions. Sequencing was performed on an ABI 3730 capillary sequencer using a 1/20 dilution of ABI Prism BigDye Terminator v3.1 sequencing chemistry after hot-start (96 °C for 30 s) thermocycling (92 °C for 5 s, 50 °C for 5 s, 60 °C for 120 s×44 cycles). Before injection, products were purified on DyeEx plates (Qiagen). PCR and RT-PCR fragments were sequenced directly with the same primers as in the PCR reaction, while M13 primers (M13-F, TGTAAAACGACGGCCAGT; M13-R, CAGGAAACAGCTATGACC) were used to sequence the cloned PCR products. The obtained sequencing chromatograms were used to quantify the methylation at a given CpG as previously described [4,30,31]. The software used for the analysis of all loci described herein is freely available at www.epigenome.org. Samples and expressed alleles were genotyped by identification of annotated SNPs in the trace files.

Analysis and statistical methods

Differential methylation in direct bisulfite sequencing experiments was determined by the Wilcoxon rank sum test [32]. DNA distance matrices and radial trees were constructed with the PHYLIP software package [33]. Distances between the internal node and the pseudogene sequence $d(H\Psi,IN)$ were computed according to $d(H\Psi,IN)=[d(H,H\Psi)+d(H\Psi,C)-d(H,C)]/2$, where $d(H,H\Psi)$ is the distance between the parental gene and the pseudogene sequences, $d(H\Psi,C)$ is the distance between the pseudogene and the chimpanzee gene sequences, and $d(H,C)$ the distance between the human and the chimpanzee gene sequences, as obtained in the DNA distance matrix for each gene group. To determine whether methylation values were conserved between parental and derivative sequences, we assigned the methylation values in two tissues per pair, where differential methylation of the pseudogene was previously detected [4]. We binned the obtained methylation values to 0, 25, 50, 75, and 100%, as shown in Table 1. Pairs displaying methylation differences lower than 25% were considered unchanged.

Acknowledgments

This work was partially funded by EU Grant LSHG-CT-2004-512066 (MOLPAGE). R.C. thanks M. Riensche, C. Haefliger, and R. Wasserkort for critically reading the manuscript and for valuable suggestions.

References

[1] P.A. Jones, D. Takai, The role of DNA methylation in mammalian epigenetics, Science 293 (2001) 1068–1070.
[2] S.B. Baylin, J.E. Ohm, Epigenetic gene silencing in cancer—a mechanism for early oncogenic pathway addiction? Nat. Rev. Cancer 6 (2006) 107–116.
[3] S.B. Baylin, DNA methylation and gene silencing in cancer, Nat. Clin. Pract. Oncol. 2 (Suppl. 1) (2005) S4–S11.
[4] F. Eckhardt, et al., DNA methylation profiling of human chromosomes 6, 20 and 22, Nat. Genet. 38 (2006) 1378–1385.
[5] M. Bibikova, et al., High-throughput DNA methylation profiling using universal bead arrays, Genome Res. 16 (2006) 383–393.
[6] B. Khulan, et al., Comparative isoschizomer profiling of cytosine methylation: the HELP assay, Genome Res. 16 (2006) 1046–1055.
[7] F. Song, et al., Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 3336–3341.
[8] C. Grunau, W. Hindermann, A. Rosenthal, Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes, Hum. Mol. Genet. 9 (2000) 2651–2663.
[9] M. Lynch, J.S. Conery, The evolutionary fate and consequences of duplicate genes, Science 290 (2000) 1151–1155.
[10] S. Ohno, Evolution by Gene Duplication, Springer-Verlag, Berlin, 1970.
[11] I. D 'Errico, G. Gadaleta, C. Saccone, Pseudogenes in metazoa: origin and features, Brief. Funct. Genomics Proteomics 3 (2004) 157–167.
[12] L.A. Naiche, Z. Harrelson, R.G. Kelly, V.E. Papaioannou, T-box genes in vertebrate development, Annu. Rev. Genet. 39 (2005) 219–239.
[13] S.L. Frank, I. Klisak, R.S. Sparkes, A.J. Lusis, A gene homologous to plasminogen located on human chromosome 2q11–p11, Genomics 4 (1989) 449–451.
[14] V.O. Lewis, et al., Homologous plasminogen N-terminal and plasminogen-related gene A and B peptides: characterization of cDNAs and recombinant fusion proteins, Eur. J. Biochem. 259 (1999) 618–625.
[15] T.E. Petersen, M.R. Martzen, A. Ichinose, E.W. Davie, Characterization of the gene for human plasminogen, a key proenzyme in the fibrinolytic system, J. Biol. Chem. 265 (1990) 6104–6111.
[16] J.L. Ashurst, et al., The Vertebrate Genome Annotation (Vega) database, Nucleic Acids Res. 33 (2005) D459–D465.
[17] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: H.N. Munro (Ed.), Mammalian Protein Metabolism, Academic Press, New York, 1969, pp. 21–132.
[18] M.E. Steiper, N.M. Young, T.Y. Sukarna, Genomic data support the hominid slowdown and an Early Oligocene estimate for the hominid–cercopithecoid divergence, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 17021–17026.
[19] G.V. Glazko, E.V. Koonin, I.B. Rogozin, Molecular dating: ape bones agree with chicken entrails, Trends Genet. 21 (2005) 89–92.
[20] L. Tsyba, A.V. Rynditch, E. Boeri, K. Jabbari, G. Bernardi, Distribution of HIV-1 in the genomes of AIDS patients, Cell. Mol. Life Sci. 61 (2004) 721–726.
[21] E.S. Balakirev, F.J. Ayala, Pseudogenes: are they "junk" or functional DNA? Annu, Rev. Genet. 37 (2003) 123–151.
[22] L. Duret, et al., The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene, Science 312 (2006) 1653–1655.
[23] S.A. Korneev, J.H. Park, M. O, 'Shea, Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene, J. Neurosci. 19 (1999) 7711–7720.
[24] S. Hirotsune, et al., An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene, Nature 423 (2003) 91–96.

[25] T.A. Gray, A. Wilson, P.J. Fortin, R.D. Nicholls, The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 12039–12044.
[26] S.N. Rodin, A.D. Riggs, Epigenetic silencing may aid evolution by gene duplication, J. Mol. Evol. 56 (2003) 718–729.
[27] C. Ivascu, R. Wasserkort, R. Lesche, J. Dong, H. Stein, et al., DNA methylation profiling of transcription factor genes in normal lymphocyte development and lymphomas, Int. J. Biochem. Cell Biol. 39 (2007) 1523–1538.
[28] K. Berlin, M. Ballhause, K. Cardon, Improved bisulfite conversion of DNA, Patent (2005) PCT/WO/2005/038051.
[29] S. Rozen, H.J. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, in: S. Krawetz, S. Misener (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology, Humana Press, Totowa, NJ, 2000, pp. 365–386.
[30] V.K. Rakyan, et al., DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project, PLoS Biol 2 (2004) e405.
[31] J. Lewin, A.O. Schmitt, P. Adorjan, T. Hildmann, C. Piepenbrock, Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates, Bioinformatics 20 (2004) 3005–3012.
[32] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.
[33] J. Felsenstein, PHYLIP—Phylogeny Inference Package (version 3.2), Cladistics 5 (1989) 164–166.
[34] V.E. Papaioannou, S.N. Goldin, Introduction to the T-box genes and their roles in developmental signaling pathways, in: C.J. Epstein, R.P. Erickson, A. Wynshaw-Boris (Eds.), Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis. Oxford Monogr. Med. Genet. No. 49, Oxford Univ. Press, 2003, pp. 686–698.